کاربرد فناوری ریز آرایه (Microarray) در تشخیص عوامل بیماری های عفونی و فرایند آنالیز داده های ریزآرایه

نظری خدیجه [۱]، دکتر کرمی علی [۲]، دکتر مهدوی امیری نظام الدین [۳]**

۱) کارشناس ارشد علوم کامپیوتر، دانشگاه صنعتی شریف ۲) دانشیار مرکز تحقیقات بیولوژی مولکولی، دانشگاه علوم پزشکی بقیه ا.. (عج) ۳) استاد دانشکده ریاضی، دانشگاه صنعتی شریف

چکیده

فناوری ریزآرایه یا میکروآری تلفیقی از توانمندی های علوم مختلف مانند بیولوژی مولکولی، میکروالکترونیک، میکروفلوئیدیک و بیوانفورماتیک است. با استفاده از این فناوری می‌توان هم‌زمان هزاران هدف ژنتیکی و یا پروتئینی را در یک مجموعه‌ی کوچک بررسی کرد.

ریزآرایه‌های DNA عبارتند از مجموعه‌ای از نقاط میکروسکوپی DNA که به سطح جامدی مثل شیشه، پلاستیک یا تراشه‌ی سیلیکون متصل شده‌اند و یک آرایه را تشکیل می‌دهند. قطعات DNA ثابت شده به عنوان کاوشگر در نظر گرفته می‌شوند. در یک آزمایش می‌توان از هزاران کاوشگر استفاده کرد. بنابراین هر ریزآرایه شامل همین تعداد آزمون ژنتیکی است که برای همه‌ی آن‌ها آزمایش به‌صورت موازی انجام می‌شود. با این توانایی ریزآرایه‌ها به بسیاری از بررسی‌های زیست شناختی سرعت بخشیده اند.

مهم‌ترین مرحله در این فناوری، تجزیه و تحلیل داده‌های انبوهی است که حاصل می‌شود و به ابزاری بیوانفورماتیکی که بتواند این داده‌ها را با درجه‌ی اطمینان بالا تحلیل کند، نیاز است.

بیماری های عفونی از آغاز زندگی بشری همواره همراه انسان بوده و باعث مشکلات اساسی شده‌اند. یکی از کاربردهای مهم فناوری ریزآرایه، امکان آزمودن وجود هزاران میکروارگانیسم در نمونه‌های محیطی یا بالینی تنها در یک آزمایش و در زمان کوتاهی است که منجر به تشخیص به‌موقع عامل عفونی و بیماری‌زا می‌شود و به‌این ترتیب گام مهمی در درمان بیماری برداشته‌ایم. همان‌طور که ذکر شد، مهم‌ترین مرحله در فناوری ریزآرایه تجزیه و تحلیل داده‌هاست. ما در این‌جا الگوریتم E-Predict و بسته ی نرم افزاریDetectiV را که بر پایه ی تعیین گونه‌های پاتوژن در ریزآرایه است، تشریح می‌کنیم. در ادامه، کاربرد این دو روش در کشف عامل بیماری‌زا در مجموعه داده‌ی بزرگی که به صورت عمومی در دسترس است، بیان می‌کنیم و نشان می‌دهیم که DetectiV بهتر از E-Predict عمل می‌کند.

واژه های کلیدی : ۱- ریزآرایه ۲- داده‌های ریزآرایه ۳- تحلیل داده‌های ریزآرایه ۴- بیماری‌های عفونی

مقدمه

روش‌های قدیمی بررسی مطالعه و تحلیل بیان ژن از جمله PCR و RT-PCR، نورترن بلات و ساترن بلات و تعیین ردیف ژن از بهترین روش‌ها برای آنالیز تعداد محدودی ژن و نمونه در یک زمان است.(۱) اما فناوری ریزآرایه قادر است هم‌زمان هزاران ژن یا پروتئین را مورد بررسی قرار دهد. نوع ریزآرایه بستگی به موادی دارد که بر روی اسلاید قرار می گیرند: اگر DNA باشد به آن ریزآرایه‌ی DNA، اگر RNA باشد ریزآرایه‌ی RNA، اگر پروتئین باشد، ریزآرایه‌ی پروتئین و اگر ماده‌ی به دست آمده از بافت ویژه ای باشد، ریزآرایه‌ی بافتی نامیده می‌شود.

مراحل آزمایشگاهی فناوری ریزآرایه شامل ساخت ریزآرایه، استخراج و نشان‌دار کردن نمونه با استفاده از رنگ‌های فلورسانتی، هیبرید کردن نمونه با ریزآرایه و شستشوی ریزآرایه است. در نهایت با استفاده از اسکنر شدت نور پروب‌های موجود در ریزآرایه اندازه‌گیری و کمی می شود و نوبت به مرحله‌ی تحلیل داده‌ها می رسد.(۲)

ریزآرایه‌های DNA از ابزارهای مفید و نوظهور در متاژنومیک هستند. این فناوری در آشکارسازی عوامل عفونی مختلف شامل انواع باکتری‌ها، ویروس‌ها، انگل‌ها، قارچ‌ها و همچنین کشف پاتوژن‌های انسانی نوظهور، بسیار موفق عمل کرده است.(۷-۳)

شاید مهم ترین کاربرد فناوری ریزآرایه در بیماری های عفونی، توانایی تشخیص سریع و هم‌زمان چند هدف و همچنین تفکیک ژنوتایپ پاتوژن‌ها می باشد است. به دلیل آن که پاتوژن‌ها ترکیبات ژنتیکی مجزا دارند و ریزآرایه‌ها قادرند که توالی همه‌ی ژن‌ها را هم‌زمان بررسی کنند، این فناوری یک ابزار ایده‌آل برای این هدف است.(۸،۹)

بسیاری از ریزآرایه ها با هدف مشخص کردن طیفی از میکروارگانیسم‌های موجود در یک نمونه طراحی شده اند و عملاً کاربرد آن‌ها در آشکارسازی ویروس‌های بیماری‌زا تایید شده است.(۱۰) شناسایی انواع عوامل تهدید بیوتروریستی(۱۱) و بررسی کیفیت آب و مواد غذایی(۱۲) جهت تشخیص انواع عوامل بیماری‌زا از دیگر کاربردهای ریز آرایه است.

به هرحال آنالیز داده‌های ریزآرایه در بیماری‌های عفونی از آن جهت که هر نمونه‌ی اسیدنوکلوئیکی مورد آزمون معمولاً شامل مخلوطی از DNA و RNA از ارگانیسم‌های مختلف، میزبان و آلودگی‌های مختلف است و همه‌ی این‌ها ممکن است در نتیجه‌ی آزمایش اثرگذار باشند، کار آسانی نیست. علاوه بر این امکان دارد که با وجود چندین گونه‌ی پاتوژن و حتی به هم وابسته، عمل هیبریداسیون پیچیده‌تر شود.

مواد و روش‌ها

در این بررسی دو روش E-predict و DetectiV جهت تحلیل داده‌های ریزآرایه‌ی هدف، استفاده شده است.

E-predict

یکی از روش‌های تحلیل داده‌های ریز آرایه استفاده از الگوریتم E-Predict است.(۱۳)

اساس این روش به منظور شناسایی گونه‌ها، الگوهای هیبریداسیون مشاهده شده‌ی ریزآرایه است. در این الگوریتم، ابتدا پروفایل‌های انرژی هیبریداسیون برای هر ژنوم ویروس مرجع به صورت کامل توالی یابی شده‌ی موجود در GenBank (شامل ۱۲۲۹ ویروس مجزا) محاسبه می‌شود.

انتظار می‌رود که همه‌ی الیگونوکلئوتیدهای ریزآرایه با ژنوم ویروس‌های معین که با استفاده از BLAST[1] (۱۴) مشخص شده‌اند، هیبرید شوند. سپس انرژی آزاد هیبریداسیون (ΔG) برای هر همترازی با استفاده از روش نزدیکترین همسایه محاسبه می‌شود. برای الیگوهایی که در تولید خروجی BLAST ناموفق هستند، انرژی هیبریداسیون صفر در نظر گرفته می‌شود. بنابراین یک پروفایل انرژی نظری مشخص شامل انرژی‌های هیبریداسیون غیر صفری است که برای زیرمجموعه ای از الیگوها که یک همترازی BLAST مطابق با ژنوم تولید می‌کنند، محاسبه می‌شود. جمعاً، پروفایل‌های انرژی از همه‌ی ویروس‌ها تشکیل

[1]NCBI BLAST [http://www.ncbi.nlm.nih.gov/BLAST]

یک ماتریس انرژی می‌دهند. در این ماتریس، هرسطر مطابق با یک گونه‌ی ویروس و هر ستون مطابق با یک الیگو از یک زیرآرایه است.

در مرحله‌ی بعد بردار شدت الیگوها نرمال می‌شود و با هر پروفایل نرمال شده در ماتریس انرژی با استفاده از یک معیار مشابهت مقایسه می‌شود. نتیجه‌ی این عملکرد یک بردار از امتیازات مشابهت و به صورت خام است. هر عنصر در این بردار، مشابهت بین الگوهای مشاهده شده و یکی از پروفایل‌های پیش بینی شده برای یک گونه‌ی موجود در ماتریس انرژی را بیان می‌کند.

در روش E-Predict مقدار p همراه با امتیاز مشابهت جواب را مشخص می‌کند.

با استفاده از E-Predict در تعدادی از وضعیت‌ها نتایج مفیدی حاصل شده است. در هر صورت، در حال حاضر E-Predict شامل ابزاری برای تصویرسازی نیست و احتیاج به بهنگام سازی و محاسبات پرهزینه قبل از استفاده‌ی آن برای آرایه‌های جدید دارد. همچنین E-Predict تنها در سیستم عامل Unix و Linux قابل استفاده است.

DetectiV

DetectiV به عنوان بسته ای برای نرم افزار آماری R شامل توابعی برای تصویرسازی، نرمال سازی و آزمون معنی‌داری از داده های ریزآرایه‌ی تشخیص پاتوژن است. R یک نرم افزار آماری رایگان است و در سیستم عامل‌های Windows، Unix، Linux، Mac قابل استفاده است.(۱۵) از آن‌جا که DetectiV در نرم افزار R ایجاد می‌شود، به آسانی با بسیاری از بسته هایی که برای تحلیل ریزآرایه در دسترس هستند، ادغام می‌شود.(۱۶)

مجموعه داده‌ی اصلی یک ماتریس از داده‌هاست که سطرهای آن نمایش دهنده‌ی پروب‌ها و ستون‌ها نمایش دهنده‌ی اندازه‌گیری‌ها از ریزآرایه هاست. این داده ها به آسانی از ساختار داده هایی که به وسیله‌ی limma به‌دست آمده اند، ایجاد می‌شوند.(۱۷) limma شامل توابعی برای خواندن فرمت‌های خروجی اکثر اسکنرهاست. affy داری توابعی برای خواندن داده‌های افی متریکس است. (۱۸)

روی پروب‌های تکراری میانگین می گیریم و در مرحله‌ی بعدی عمل نرمال‌سازی را با استفاده از تقسیم مقادیر متناظر با هر پروب بر مقدار میانه‌ی آرایه‌ی متناظر و سپس لگاریتم گیری در مبنای ۲ انجام می‌دهیم. به این ترتیب داده‌ها نرمال و میانگین آن‌ها صفر است. (البته فرض بر این است که بیشتر پروب ها با چیزی هیبرید نمی شوند).

بعد از آن با استفاده از آزمون t داده‌ها براساس گونه‌ی ویروس گروه‌بندی می‌شوند. داده‌هایی در مراحل بعدی استفاده می‌شوند که لگاریتم میانگین در مبنای ۲ آن‌ها، بزرگتر یا مساوی ۱ باشد. این اعداد بر اساس مقدار p مرتب می‌شوند.

در این مبحث داده‌های استفاده شده از GEO با عدد دست‌یابیGSE2228 پیاده شده‌است.(۱۹) پلت فرم آرایه برای این داده‌ها، GEO[1] با عدد دست‌یابی GPL1834 است و شامل بیش از ۱۱۰۰۰ الیگوست که نشان دهنده‌ی ۱۰۰۰ گونه‌ی ویروس و باکتری است.

نتایج و بحث

نتایج مورد انتظار و صحیح از قبل مشخص شده‌اند در صورتی نتایج به دست آمده از تحلیل داده‌ها صحیح فرض می‌شود که یکی از دو مورد زیر اتفاق بیافتد:

۱) جواب حاصل از تحلیل با عامل بیماریزای از قبل مشخص مشابه باشد.

۲) اگر در آرایه عامل بیماری‌زای موردنظر وجود نداشته باشد، آنگاه نتیجه‌ی حاصل بسیار مشابه و وابسته به ویروس مورد نظر باشد.

[1] Gene Expression Omnibus[http://ncbi.nml.nih.gov/GEO

DetectiV در ۵۵ آرایه از ۵۶ آرایه‌ی موجود منجر به نتیجه‌ی صحیح می شود. در مقایسه، روش E-Predict در ۵۳ آرایه از ۵۶ آرایه نتیجه‌ی صحیح می دهد.

DetectiV از تکنیک های خوب و ساده‌ی تصویرسازی و روش‌های آماری مناسب برای تحلیل داده‌های ریزآرایه‌های آشکارسازی پاتوژن استفاده می‌کند.

ما در اینجا بین مفاهیم آماری و زیستی تفاوت قایل می‌شویم. در روش DetectiV کمترین مقدارp در ترکیب با بزرگترین مقدار میانگین، و در روش E-Predict کمترین مقدار p همراه با بزرگترین مقدار مشابهت جواب را مشخص می کند.

در بسیاری از حالات با استفاده از قوانین خودکار می توانیم جواب درست را نتیجه بگیریم. به هر حال، به ناچار بعضی حالات وجود دارند که نیازمند دخالت انسان است.

نتایج کاربرد DetectiV روی مجموعه داده‌ی GSE8746 با جواب درست در هر ۱۲ آرایه مشوق بسیار خوبی برای ماست. به‌ویژه، توانایی آرایه و DetectiV در تمیز دادن بین نه‌تنها گونه های ویروس بلکه زیرنوع‌های FMDV بسیار جذاب و درخور توجه است و نمایانگر قدرت آن است.

نتایج کاربرد DetectiV روی مجموعه داده‌ی سارس نیز مشوق و امیدوار کننده است. در این‌جا، الیگوهایی که مخصوص ویروس سارس طراحی شده باشند، روی آرایه وجود ندارند.

این نکته شایان ذکر است که برای کاربرد DetectiV در مجموعه داده‌ی دیگری که از آرایه‌های کاملاً متفاوت با مجموعه داده‌ی اول استفاده می کند، تنها نیاز است که کاربر عدد دست‌یابی GEO و تعداد آرایه‌های داخل مجموعه داده را تغییر دهد. این را می توان با E-Predict مقایسه کرد که نیازمند محاسبه‌ی ماتریس مشابهت پیچیده و بزرگ و همچنین بهینه‌سازی چندین پارامتر است.

References

1. Brown T.A.N: Gene cloning and DNA analysis an introduction. *Blackwell Science 2006.*
2. Wentian L, Yaning Y : Introduction to microarray analysis. *BMC proceedings 2007*
3. Riesenfeld CS, Schloss PD, Handelsman J: Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet 2004.*
4. Lapa S, Mikheev M, Shchelkunov S, Mikhailovich V, Sobolev A, Blinov V, Babkin I, Guskov A, Sokunova E, Zasedatelev A, et al.:Species level identification of orthopoxviruses with an oligonucleotide microchip. *J clin Microbiol 2002.*
5. Boonham N, Walsh K, Smith P, Madagan K, Graham I, Barker I: Detection of potato viruses using microarray technology:towards a generic method for plant viral disease diagnosis. *J virol Methods 2003*.
6. Song Y, Dai E, Wang J, Liu H, Zhai J,Chen C, Du Z, Guo Z, Yang R: Genotyping of hepatitis B virus HBV by oligonucleotides microarray. *Mol Cell Probes 2006.*
7. Perrin A, Duracher D, Perret M, Cleuziat P, Mandrand B:A combined oligonucleotide and protein microarray for the codetection of nucleic acids and antibodies associated with human immunodeficiency virus, hepatitis B virus and hepatitis C virus infections. *Anol Biochem 2003.*
8. Mezzasoma L, Crisanti A, Rossi R: Antigen Microarrays for Serodiagnosis of Infectious Diseases. *Clinical Chemistry 2002.*
9. Mohammed Z, Souna E, Anthony T: Principles of Bacterial Detection: Biosensors, Recognition Receptors and Microsystems.*Springer-Science and Business Media 2008.*
10. Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL: Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA 2002.*
11. Sergeev N, Distler M, Courtney S, Al-Khaldi SF, Volokhov D,Chizhikov V, Rasooly A: Multipathogen oligonucleotide microarray for environmental and biodefense applications. *Biosens Bioelectron 2005.*
12. Lemarchand K, Masson L, Brousseau R: Molecular biology and DNA microarray technology for microbial quality monitoring of water. *Crit Rev Microbiol 2004.*

13. Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, Wang D, DeRisi JL: E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol 2005.*
14. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic Local Alignment Search tool. *J Mol Biol 1990.*
15. The R Project for Statistical Computing [http://www.R-project.org]
16. Michael Watson, Juliet Dukes, Abu-Bakr Abu-Median, Donald P King, Paul Britton: DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data. *Genome Biol 2007.*
17. Smyth GK: Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* Edited by: Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New *York: Springer.2005*
18. Gautier L, Cope L, Bolstad BM, Irizarry RA: affy - analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics 2004.*
19. Tanya Barrett, Ron Edgar: Gene Expression Omnibus (GEO): Microarray data storage, submission, retrieval, and analysis. *Methods Enzymol 2006.*

Abstract

Microarray technology is the combination of different Sciences utilities such as Molecular Biology, Microelectronics, Microfluidics and Bioinformatics. By using this technology we can investigate, simultaneously, thousands of genetic or protein targets in a small system.

DNA microarrays consist of DNA microscopic points that are attached to a solid surface such as glass, plastic or silicon chip and formed as an array. The fixed pieces of DNA are considered as searchers. In an experiment, we can use thousands of searchers. Therefore, any microarray consists of the same number of genetic tests as the experiment performed on all of them in parallel. Whit this ability, arrays have speeded up the biological investigations. Microarray technology can be seen as a continued development of southern blotting.

However, the most important stage in this technology, analysis of data, requires reliable bioinformatics tools achieving high reliabilities.

Infectious diseases, from the beginning of human life, always have been with human beings and have caused major difficulties for them. One of the most important usages of microarray technology is the possibility of testing for the presence of thousands of micro-organism in environmental and clinical samples only in a single experiment, resulting in recognition of pathogens. Thereby, we can take an important step in curing the disease. As noted above, the most important stage in microarray technology is the data analysis. We present E-Predict algorithm and DetectiV package, for microarray based species identifications. We demonstrate the application of E-Predict and DetectiV to viral detection in a large, publicly available dataset and show that DetectiV performs better than E-Predict.

Keyword: 1-Microarray 2- Microarray Data 3- Microarray Data Analysis4- Infectious Diseases